UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING
SEMESTER II, SESSION 2024/2025

BACHELOR OF COMPUTER SCIENCE (BIOINFORMATICS)

SECB3203 PROGRAMMING FOR BIOINFORMATICS - SECTION 01

PROJECT PROPOSAL
ALZHEIMER'S DISEASE PREDICTION

GROUP MEMBERS	MATRIC NO
TAN ZHAO HONG	A23CS0188
CHIN PEI WEN	A23CS0065
KOO XUAN	A23CS0300

LECTURER'S NAME : DR. SEAH CHOON SEN

SUBMISSION DATE : 05 DECEMBER 2025

Table of Contents

1. Introduction

Alzheimer's Disease (AD) represents one of the most pressing challenges in modern healthcare, being the most prevalent form of dementia worldwide. As the global population ages, the incidence of AD continues to rise dramatically, affecting millions of individuals and their families. Early and accurate diagnosis of Alzheimer's Disease is critical for timely intervention, treatment planning, and improving patient quality of life.

This project aims to develop a robust machine learning-based classification system for Alzheimer's Disease prediction using comprehensive clinical, demographic, and lifestyle data. By leveraging advanced computational techniques and bioinformatics approaches, we seek to create an effective predictive model that can assist healthcare professionals in early AD detection and risk assessment.

The availability of large-scale healthcare datasets and advancements in machine learning algorithms present unprecedented opportunities to identify patterns and biomarkers associated with Alzheimer's Disease. Through this project, we will explore various machine learning techniques, including feature selection, data preprocessing, model development, and evaluation to create an accurate and interpretable classification system.

2. Problem Background

Alzheimer's Disease is a progressive neurodegenerative disorder characterized by cognitive decline, memory loss, and behavioral changes. According to recent studies, AD affects approximately 35.6% of diagnosed dementia cases, with the remaining 64.4% representing individuals without the disease. The challenge lies in the complexity of AD diagnosis, which traditionally relies on:

- **Clinical assessments:** Cognitive and functional tests that may be subjective
- **Medical imaging:** MRI and PET scans that require specialized equipment and expertise
- **Medical history:** Comprehensive patient background that may be incomplete
- **Laboratory tests:** Biomarker analysis that can be expensive and time-consuming

Traditional diagnostic methods often result in:

1. **Late-stage diagnosis:** Many patients are diagnosed only after significant cognitive decline
2. **Inconsistent results:** Variability between different assessment methods and practitioners
3. **Resource constraints:** Limited access to specialized diagnostic facilities, particularly in rural or underserved areas

4. **High costs:** Expensive imaging and laboratory procedures that may not be covered by insurance

The dataset we will utilize contains records from 2,149 patients with 34 comprehensive features spanning multiple domains:

- **Demographic factors:** Age, gender, ethnicity, education level
- **Lifestyle factors:** Physical activity, diet quality, sleep patterns, alcohol consumption, smoking status
- **Medical history:** Family history of AD, cardiovascular disease, diabetes, hypertension, depression
- **Clinical measurements:** BMI, blood pressure, cholesterol levels
- **Cognitive assessments:** MMSE scores, functional assessment scores, memory complaints
- **Symptoms:** Behavioral changes, confusion, disorientation, personality changes

This rich dataset provides an opportunity to develop machine learning models that can identify patterns and risk factors associated with Alzheimer's Disease, potentially enabling earlier detection and intervention.

3. Problem Statement

Current Alzheimer's Disease diagnostic approaches face several critical challenges:

1. **Delayed diagnosis:** Most patients are diagnosed in moderate to advanced stages when cognitive decline is already significant, limiting the effectiveness of available interventions
2. **Limited accessibility**: Advanced diagnostic tools such as specialized imaging (PET scans, advanced MRI) and biomarker testing are not readily available in all healthcare settings, particularly in developing regions
3. **High diagnostic costs:** Comprehensive AD assessment can be prohibitively expensive, involving multiple specialist consultations, imaging studies, and laboratory tests
4. **Subjective assessments:** Many cognitive tests rely on clinical judgment, which can vary between practitioners and may be influenced by patient demographics and socioeconomic factors
5. **Data integration challenges:** Patient information is often scattered across multiple sources (clinical notes, lab results, imaging reports), making it difficult to obtain a holistic view for accurate diagnosis
6. **Lack of predictive tools:** Current methods primarily focus on confirming diagnosis rather than predicting risk, missing opportunities for preventive interventions

4. Objectives

The primary purpose of this project is to explore how computational and bioinformatics approaches can be applied to analyse clinical datasets and support biomedical research. The project aims :

- To demonstrate practical use of programming tools for data cleaning, data interpretation, and predictive modelling.
- To provide insights into how machine-learning methods may assist in distinguishing between different stages of Alzheimer's disease.
- To cultivate understanding of how computational models contribute to addressing real healthcare challenges.

5. Scopes

The analytical process follows a structured pipeline inspired by, but not limited to, the referenced Kaggle code. The dataset will first be imported and examined using descriptive statistics to understand feature distribution and identify missing data. A preprocessing stage will then be conducted, involving tasks such as handling missing values, encoding categorical variables, and normalizing selected features where necessary. Exploratory Data Analysis (EDA) will be performed to visualize trends, detect patterns, and explore relationships between variables and disease categories. Following EDA, several machine-learning classification algorithms, such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) will be implemented to evaluate their performance in predicting Alzheimer's disease status. Model evaluation will be carried out using metrics such as accuracy, precision, recall, F1-score and confusion matrices to determine the most effective predictive approach. This methodology ensures a systematic process from data understanding to model development and performance assessment.

By completing this project, we expect to:

- Identify the most important features that influence Alzheimer's diagnosis.
- Build a machine-learning model capable of predicting disease stages.
- Understand how programming tools are used in bioinformatics research.
- Gain hands-on experience working with health-related datasets.

6. Conclusion

In summary, this project employs a computational framework to investigate Alzheimer's disease prediction using a publicly available clinical dataset. The study reflects the growing reliance on bioinformatics tools in medical research and highlights the role of machine learning in supporting diagnostic decision-making. Through systematic data analysis and model evaluation, the project aims to contribute to broader efforts in understanding and detecting neurodegenerative diseases at earlier stages.

7. Reference

- Rabie El Kharoua, "🧠 Alzheimer's Disease Dataset 🧠," *Kaggle.com*, 2024, doi: https://doi.org/10.34740/KAGGLE/DSV/8668279%7D,/n/tpublisher=%7BKaggle%7D,/n/tauthor=%7BRabie.

- adhamtarek147, "Alzheimer's Disease Prediction 🧠🔍📈," *Kaggle.com*, Jul. 09, 2024. https://www.kaggle.com/code/adhamtarek147/alzheimer-s-disease-prediction

- Wikipedia Contributors, "Alzheimer's disease," *Wikipedia*, Jan. 25, 2019. https://en.wikipedia.org/wiki/Alzheimer%27s_disease

- Hala Alshamlan, Arwa Alwassel, Atheer Banafa, and Layan Alsaleem, "Improving Alzheimer's Disease Prediction with Different Machine Learning Approaches and Feature Selection Techniques," *Diagnostics*, vol. 14, no. 19, pp. 2237–2237, Oct. 2024, doi: https://doi.org/10.3390/diagnostics14192237.